CAPSTONE PROJECT

Validation, Monitoring, and Governance

For

Enhancing Loan Approval with Predictive Modeling

IBUKUNOLUWA OLUKOKO-301259629

Validation, Monitoring, and Governance

1.1 Introduction

In a world where data's a crucial factor machine learning models such as our loan prediction model have become instrumental in shaping business choices. These models, including our regression tree-based loan prediction model, hold potential to transform lending practices streamline operations and improve customer satisfaction. Nonetheless, the changing nature of data, market conditions and regulatory demands call for a validation, monitoring and governance framework to guarantee the model's dependability and efficacy in the long run.

This section outlines our approach to validation, monitoring, and governance of the regression tree loan prediction model. Through systematic validation practices and ongoing monitoring, we ensure that the model maintains its predictive accuracy and alignment with business objectives. By adhering to these practices, we mitigate risks associated with model deterioration and maintain a high level of confidence in its performance.

To achieve these goals, we focus on variable level monitoring, model health assessment, and risk tiering strategies. By closely observing feature behavior, tracking model performance metrics such as accuracy and F1 score, and implementing responsive actions, we ensure the model remains aligned with business requirements. Through continuous monitoring and validation, we can promptly identify and address potential issues, adapting to changing circumstances effectively.

In the following sections, we will provide an explanation of our validation and governance procedures. This will include how we keep track of the model variables, assess its performance, and address any potential risks. By outlining these practices, we hope to demonstrate our dedication to maintaining an efficient loan prediction model. Our ultimate goal is to ensure that this model aids in making lending decisions while supporting our organization's achievements in a changing data-centric landscape.

1.2 Variable Level Monitoring

This involves the continuous observation and assessment of individual variables within a dataset. This process enables the identification of any anomalies, shifts, or changes in the distribution, behavior, or characteristics of these variables over time. By closely monitoring variables, organizations can promptly detect deviations, assess their impact, and take necessary actions to maintain data quality and model

performance. This practice contributes to the overall stability, reliability, and effectiveness of data-driven decision-making processes.

1.2.1 Model Build Variable Level Statistics

To fully grasp the governance and validation procedures it is essential to have an understanding of the attributes associated with the variables utilized in our model. Descriptive statistics, such as median and standard deviation are employed to summarize features. In order to assess distribution changes over time for features we employ bar charts. This allows us to detect any variations in feature distributions that may impact the performance of the model.

Given that the variables in the data are numerical with Categorical meaning described as,

1. APP_ID: Application id's uniquely identifying each loan applicant.

2. CIBIL_SCORE_VALUE: Cibil score, represented as three digits, indicates the creditworthiness of the applicant, with values 0=bad, 1=ok, and 2=good.

3. NEW_CUST: New To Credit score helps banks rate new borrowers, indicated as "Y" or "N."

4. CUS_CATGCODE: Customer category code categorizes applicants as existing or new customers.

5. EMPLOYMENT_TYPE: Indicates the type of employment, with 1 representing salaried and 0 representing self-employed applicants.

6. Age: Represents the age of the loan applicant.

7. SEX: Gender of the person, with "F" for female and "M" for male.

8. NO_OF_DEPENDENTS: Specifies the number of dependents of the applicant.

9. MARITAL_Status: Marital status of the applicant, with 1 representing married and 0 representing not married.

10. EDU_QUA: Educational qualifications of the applicants, where 1 indicates educated and 0 indicates not-educated.

11. P_RESTYPE: Type of residence of the applicant, with 1 indicating own residence and 0 indicating rented residence.

12. P_CATEGORY: Type of house the applicant stays in.

13. EMPLOYEE_TYPE: Represents the type of employment, where 0=Private, 1=Temporary, and 2=Government.

14. MON_IN_OCC: Indicates the number of months the applicant has been working in the present employment.

15. INCOME_EXP_GMI: Income Expenses based rating.

16. ASSET_LOAN_RATIO: Pre-calculation based asset & loan ratio.

17. TENURE: Loan tenure ranges from 12 to 48 months.

18. Status: Represents loan approval (1) or rejection (0).

Only those with proper integer meaning can have a statistical description run on then, which will then result in

Descriptive Statistics for Selected Columns:

Descriptive Statistics for Selected Columns:

	AGE	MON_IN_OCC	INCOM_EXP_GMI	LTV	TENURE	NO_OF_DEPENDENTS
count	13299.000000	13299.000000	13299.000000	13299.000000	13299.000000	13299.000000
mean	32.473870	71.946161	0.667644	0.781083	24.792240	1.536281
std	8.804317	65.114130	0.762374	0.108869	7.501085	0.971671
min	18.000000	1.000000	0.000000	0.500000	12.000000	0.000000
25%	25.000000	24.000000	0.000000	0.719710	18.000000	1.000000
50%	31.000000	48.000000	0.000000	0.807537	24.000000	2.000000
75%	38.000000	100.000000	1.000000	0.849177	36.000000	2.000000
max	55.000000	240.000000	2.000000	1.000000	36.000000	3.000000

1.2.2 Acceptable Ranges, Caps & Floors

Each feature has an established acceptable range defined by its minimum and maximum values. If future values fall outside this range and are flagged as outliers, caps and floors are applied to ensure values remain within an acceptable range. This approach maintains the integrity of the data and prevents extreme values from impacting the model's behavior.

1.2.3 No Missing Values

All the features in the dataset have complete and accurate data, and there are no missing values. This ensures that our model is trained on comprehensive data, minimizing any potential impacts related to missing values.

1.2.1 Variable Drift Monitoring

The predictive models developed for this project provide an understanding of the factors that influence decisions about approving loans. The insights they offer allow the banking organization to address risks proactively and make informed choices. However, it is crucial to recognize that the variables used in training these models may undergo changes or shifts in significance over time due to factors.

The impact of data drift goes beyond prediction accuracy; it also includes the relevance of the features used in the analysis. For example, changes in the job market or economic conditions can cause shifts in the distribution of variables like Employment Type. Similarly, if there are changes in the landscape of loan applicants Age might become less relevant for predicting loan approvals.

This highlights the importance of monitoring both significance and relevance of features included in our models. If customer profiles or industry trends change within the banking organization certain features might lose their power. To ensure that our model performs consistently, and reliably effective model governance requires us to address these shifts.

Ultimately our goal is to maintain and uphold our model's capabilities while generating actionable insights for making loan approval decisions.

It is crucial to monitor data changes and adjust the model accordingly to avoid any decline in its performance. This ensures that the model continues to aid in the decision-making processes of banking organizations.

1.2.2 Tolerance for Drift of Each Variable

Drift tolerance for the more important features is set at 0.8173, which is 5% of the mean of the features used in model training. For the less important features, the drift tolerance is set at 2.1698, which is 12% of the mean of the features used in model training. When the drift of a feature exceeds its respective threshold, instead of rebuilding the models directly, model health and stability will be first assessed by looking into the AUC-ROC and F1-score. Subsequently, appropriate actions will be taken based on the risk tiering.

1.3 Model Monitoring and Health

The robustness and reliability of the models are continuously monitored through a vigilant process of model health assessment. Key features are tracked for potential drift, with predefined drift tolerance

thresholds established based on comprehensive statistical analysis. Should any feature's drift surpass its respective threshold, a meticulous evaluation of model performance metrics, including AUC-ROC and F1-score, ensues. This rigorous monitoring approach ensures the models' stability and informs appropriate actions aligned with risk tiering strategies, thereby safeguarding the efficacy and accuracy of the predictive models.

1.3.1 Initial Model Fit Statistics

Upon building the model, we evaluate its initial performance using key metrics such as accuracy, F1 score, precision, recall, Jaccard score, and log loss. These metrics provide a baseline understanding of the model's effectiveness and guide subsequent validation and monitoring steps.

1.3.2 AUC-ROC and F1 Score Evaluation

The Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve and the F1 score are chosen as essential performance indicators. The AUC-ROC evaluates the model's overall discriminatory power, while the F1 score balances precision and recall. These metrics enable us to assess the model's performance comprehensively.

1.4 Risk Tiering

1.4.1 Risk Tiering Criteria

To categorize and address potential model issues effectively, we implement a risk tiering framework. This framework classifies risks based on predefined thresholds for AUC-ROC and F1 score drift. Depending on the level of drift, appropriate actions are taken to maintain model accuracy.

1.2.1 Initial Model Fit Statistics and Risk Tiering

The model risk management framework emphasizes the regular assessment of model performance through both qualitative and quantitative methods, ensuring the model's effectiveness in predicting outcomes and guiding business decisions. To comprehensively evaluate algorithmic health and monitor model stability over time, the Area Under the Curve (AUC) of the Receiver-Operator-Characteristic (ROC) curve and the F1 Score have been chosen as essential diagnostic tools and performance benchmarks.

AUC-ROC:

The selected model in this project, the Random Forest with Sampling, exhibits an AUC of 0.9918.

F1 Score:

The chosen model for this project achieves an F1 Score of 0.9515.

1.4.2 Action Steps for Each Risk Tier

Low-Risk Tier: With drift below 0.8173 for more important features and 2.1698 for less important features (both under 2%), no immediate intervention is necessary. The models remain stable and aligned with the original training data.

Moderate Risk Tier (2% - 5% drift): Drift within the range of 2% to 5% indicates potential shifts in data distribution. To optimize model performance, strategies such as hyperparameter tuning and ensemble learning methods are applied. These techniques help recalibrate the model and adapt to evolving data trends.

High-Risk Tier (6% - 10% drift): When drift reaches the range of 6% to 10%, model refitting is conducted using new data samples. This proactive measure addresses the impact of evolving data patterns, ensuring the model remains robust and accurate over time.

Unacceptable Risk Tier (above 10% drift): Drift surpassing 10% suggests significant structural changes in data. In this scenario, model rebuilding becomes essential to maintain reliable predictions. Reconstructing the model from scratch with updated data ensures its relevance and effectiveness in capturing the evolving landscape.

The tiered approach to action steps aligns with the drift tolerance thresholds, enabling the organization to maintain model health and stability while addressing data shifts appropriately.

1.5 Conclusion

The validation, monitoring, and governance processes outlined in this documentation are vital to maintaining the reliability and performance of our loan prediction model. Continuous monitoring, analysis, and appropriate actions ensure that our model continues to provide accurate predictions, aligning with business goals and regulatory standards.

Through these processes, we not only reduce potential risks but also enhance our model's stability, accuracy, and predictive power in a dynamic business environment.